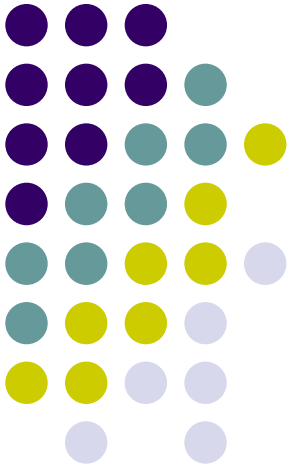
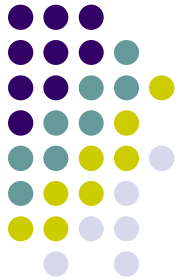
Компьютерная лингвистика: от текста к знаниям

Вводная лекция о современной компьютерной лингвистике и её приложениях в реальной жизни



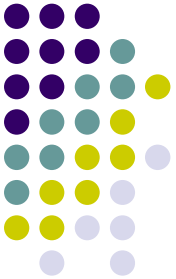
Даниил Скатов, ООО «Диктум»
г.Н.Новгород, ННГУ, 31 окт. 2011

Суть предмета



- Англ. оригинал — *computational linguistics*
- Смысл широкий: (1) обработка естественного языка, анализ и синтез текста; (2) распознавание символов; (3) распознавание и синтез речи. Обычно относят к (1).
- Если детальнее по (1): information retrieval; работа с источниками лингвистических данных (корпуса, словари); трансляторы текстов (машинный перевод, автореферирование).
- В основном будем говорить про **information retrieval**
- **Знание** — форма существования и систематизации результатов познавательной деятельности человека. Помогает людям действовать рационально. **Задача:** неструктурированная информация → структурированная информация

Наиболее яркие приложения



- **Поисковые системы**: Google и пр. — без комментариев ☺; удобный поиск в юридических документах; поисковая подсказка с исправлением опечаток для разных языков
- **Маркетинг**: оценка мнений о товарах и услугах, планирование выпуска продукции, управление репутацией персон и брендов
- **Фактографический поиск**: детальный портрет человека или предмета; прогнозирование и предотвращение событий, госбезопасность

Стандартный конвейер NLP



Токенизация → Опечатки → Морфология → Синтаксис → Постпроцесс

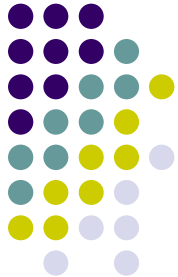
Город$_1$ *Нижний*$_2$ *Новгород*$_3$
является$_4$ *столицей*$_5$ *Поволжья*$_6$

ニージニーノブゴロドでは、
ヴォルガの首都です。

В 2014 ***г. С****очи примет гостей Олимпиады*

[Город Нижний Новгород]$_1$
является$_2$ *столицей*$_3$ *Поволжья*$_4$

Стандартный конвейер NLP



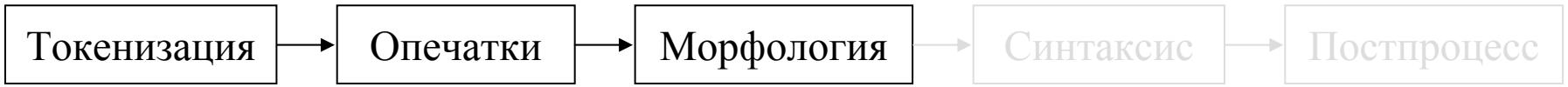


Город Нижний Новгород
являеться сталицей *Поволжья*$_6$

сталицей → столицей, станицей, старицей

являеться → является

Стандартный конвейер NLP



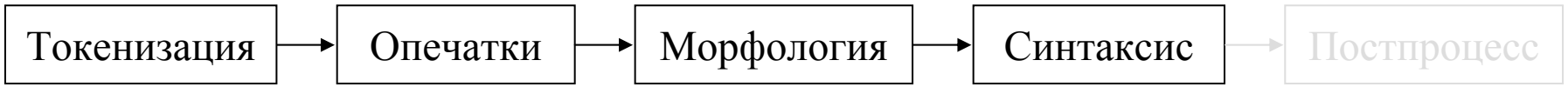
Словоформа → {мн-во г.з. + лексемы}

По сути = некоторое соотнесение
словоформы с действительностью

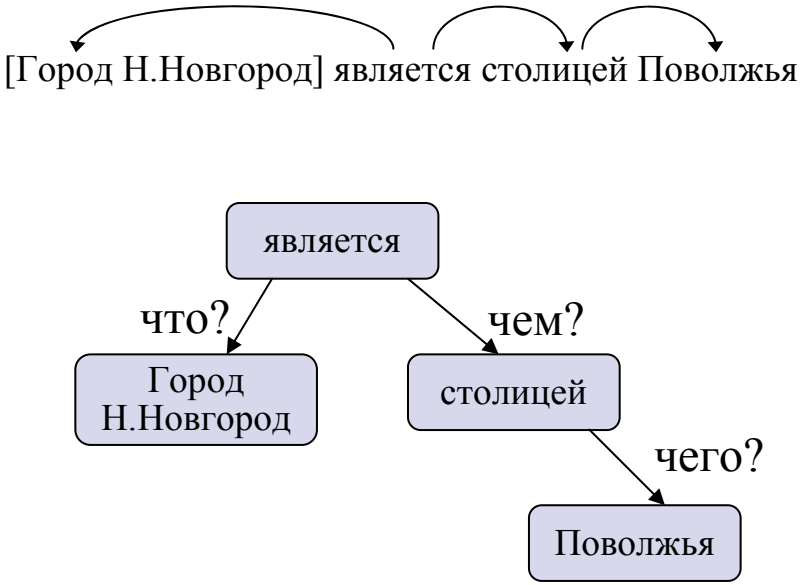
Город → (город, неодуш., ед.ч., им.п.)
Поволжья → (Поволжье, сущ., геогр., неодуш., род.п., ед.ч.),
(Поволжье, сущ., геогр., неодуш., им.п., ед.ч., редк.),
(Поволжье, сущ., геогр., неодуш., вин.п., ед.ч., редк.)
правила → (правило, сущ., ед.ч., род.п., неодуш.),
(правило, сущ., мн.ч., им.п., неодуш.),
(править, глаг., несов., прош., ед.ч., ж.р.)

Глокая куздра штеко будланула
бокра и кудрячит бокрёнка.

Стандартный конвейер NLP



По сути — сопоставление слов предложения правилам его построения



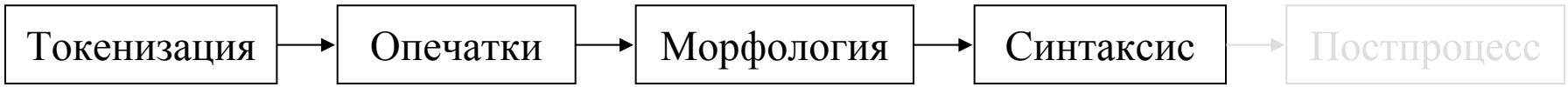
Иван идёт
Иван видит Петра
Молодой Иван видит старого Петра
Молодой Иван видит старого идущего Петра

```
S → NP VP
VP → Verb
VP → Verb NP
NP → Noun
NP → Adjective NP
NP → Preposition NP
```

```
Adjective — молодой
Adjective — старого
Adjective — идущего
Noun — Иван
Noun — Петра
Verb — видит
Verb — идёт
```

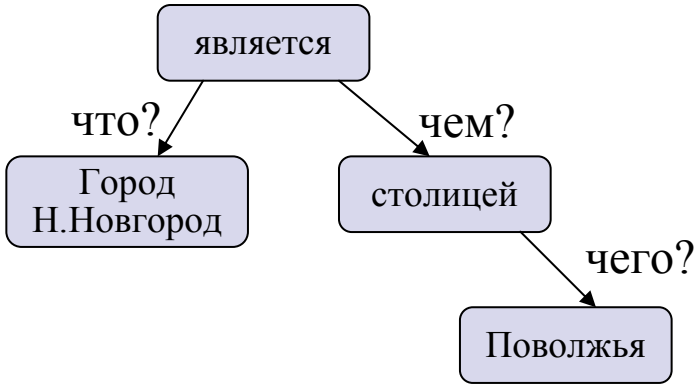
Проблемы: свободный порядок слов (особенно для русского), описание согласования (прил. + сущ. по р.,ч.,п.) → комбинаторный характер, громоздкость

Стандартный конвейер NLP



По сути — сопоставление слов
предложения правилам его построения



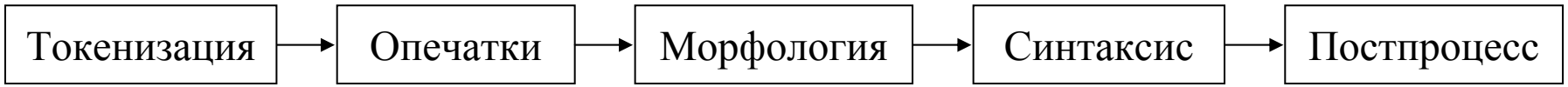




$$O(2^n) \rightarrow O(n^5) \rightarrow O(n^3)$$

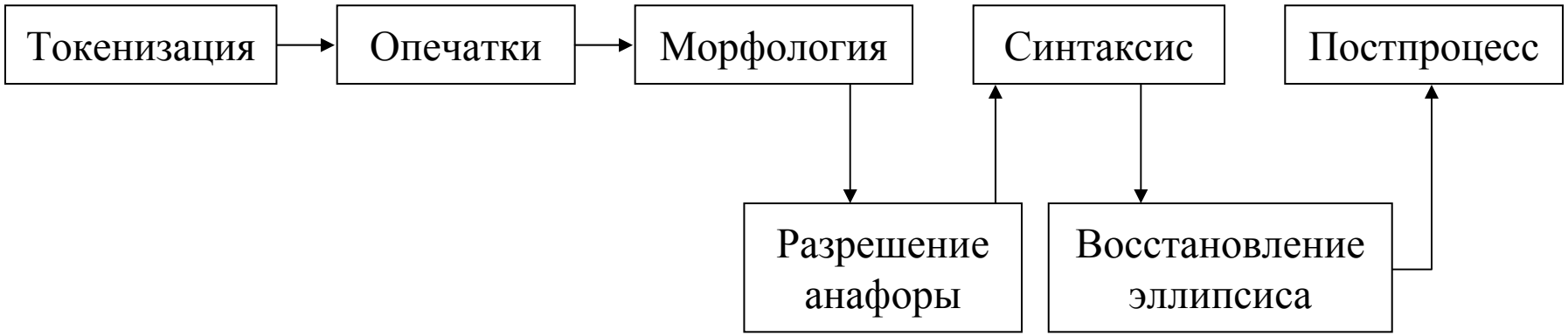
- Сдавайте мусор дворнику, который накопился
- Бытие определяет сознание
- Отдай бананы обезьянам — они голодные
- Отдай бананы обезьянам — они перезрелые
- Эти типы стали есть на складе

Стандартный конвейер NLP



Обработка (обход, сопоставление с шаблонами и т.д.)
полученных деревьев и именованных сущностей
с целью добычи знаний — оценка тональности,
установление фактов, выполнение логического вывода и т.п.

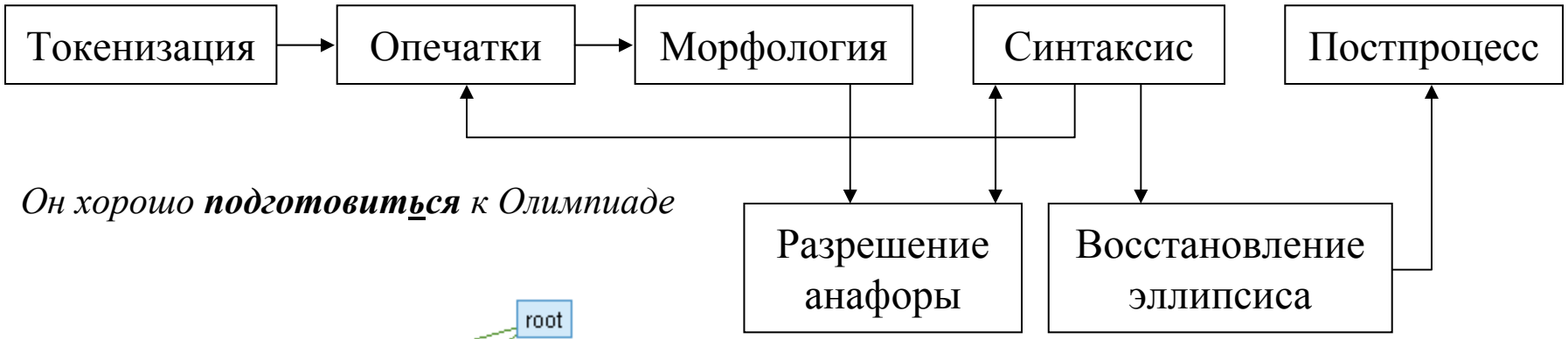
Стандартный конвейер NLP



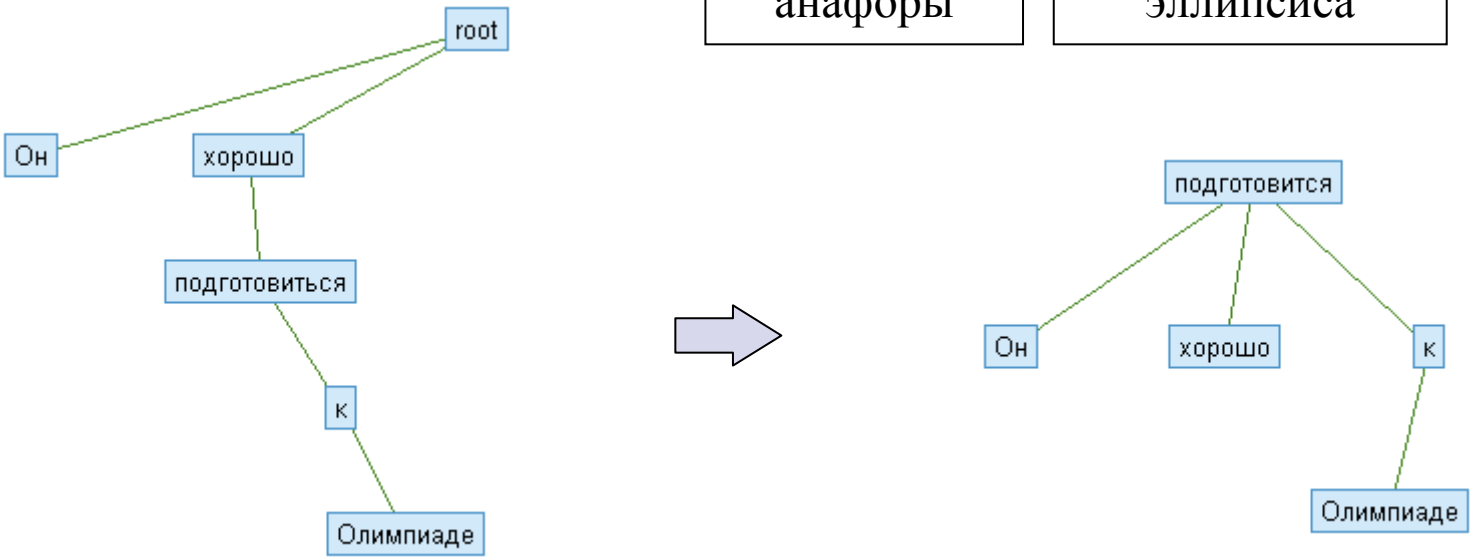
Купил телефон для бизнеса. У него ***{телефона, бизнеса}*** *чёткий экран, но он* ***{телефон, бизнес, экран}*** *бликует*

У Siemens есть большой экран, а у Samsung — нет ***{большого экрана}***

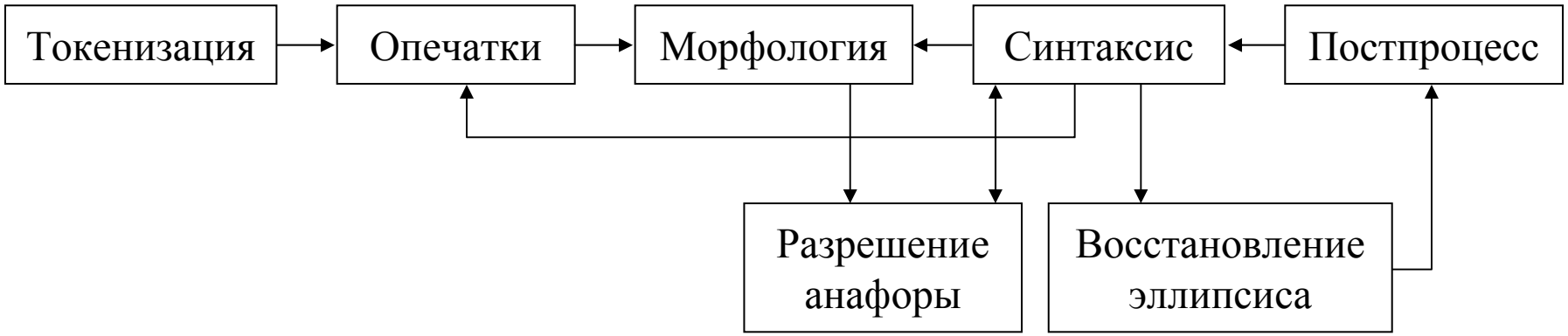
Стандартный конвейер NLP



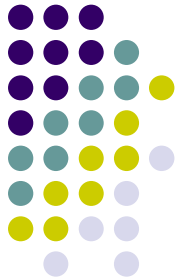
*Он хорошо **подготовиться** к Олимпиаде*



Стандартный конвейер NLP

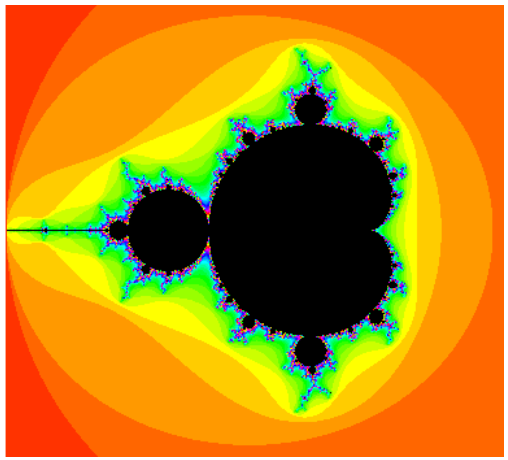


Math vs Rule-Based



«Неявный»

- Используются математические результаты — машинное обучение, статистическая механика и т.д.
- Результат сложно модифицировать, человек редко имеет возможность дать ему структурную интерпретацию

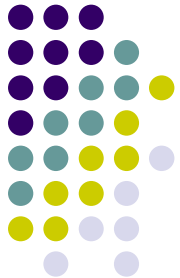


«Явный»

- Используются DSLs (domain specific languages) для описания управляющих правил
- Большой объём ручной работы
- Результаты могут получиться негибкими
- Сложно контролировать рост

$1 + 1 > 2$

Технические аспекты



- Технологии для боевого применения: десятки тысяч новых документов ежедневно, тысячи запросов в секунду
- Всё должно быть быстрым → оптимальные алгоритмы и структуры
- После предела → оптимизация кэш-промахов, компактность данных
- Локальные оптимизации, профилирование
- ОС: в основном FreeBSD и Linux
- → Языки: в основном C++
- Обеспечение компактности описаний → хорошо продуманные DSL